STIC-ILL 5359992 ~~NPL~~

QH426.G46
NO3

376463

From: Goldberg, Jeanine
Sent: Thursday, December 20, 2001 3:05 PM
To: STIC-ILL
Subject: please pull hgt1 references

1. AMERICAN JOURNAL OF MEDICAL GENETICS, (1999 Dec 15) 88 (6) 694-9.
   Journal code: 3L4; 7708900. ISSN: 0148-7299.

2. HUMAN MOLECULAR GENETICS, (2000 Jul 22) 9 (12) 1753-8.
   Journal code: BRC. ISSN: 0964-6906.

3. GENE, (2001 May 30) 270 (1-2) 69-76.
   Journal code: FOP; 7706761. ISSN: 0378-1119.

4. Genomics Vol 32 (1) pages 75-85 1996

5. Genomics Vol 25 No. 3, pages 707-715 1995. 7759106

6. ARCHIVES OF NEUROLOGY, (2001 Oct) 58 (10) 1649-53.
   Journal code: 80K; 0372436. ISSN: 0003-9942.

THANK YOU

Jeanine Enewold Goldberg
1655
CM1--12D11
Mailbox-- 12E12
306-5817

Structural Analysis of the 5′ Region of Mouse and Human Huntington Disease Genes Reveals Conservation of Putative Promoter Region and Di- and Trinucleotide Polymorphisms

BIAOYANG LIN, JAMAL NASIR, MICHAEL A. KALCHMAN, HELEN MCDONALD,
JUTTA ZEISLER, Y. PAUL GOLDBERG, AND MICHAEL R. HAYDEN[1]

Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z4

Received August 18, 1994; revised November 11, 1994

We have previously cloned and characterized the murine homologue of the Huntington disease (HD) gene and shown that it maps to mouse chromosome 5 within a region of conserved synteny with human chromosome 4p16.3. Here we present a detailed comparison of the sequence of the putative promoter and the organization of the 5′ genomic region of the murine (*Hdh*) and human HD genes encompassing the first five exons. We show that in this region these two genes share identical exon boundaries, but have different-size introns. Two dinucleotide (CT) and one trinucleotide intronic polymorphism in *Hdh* and an intronic CA polymorphism in the HD gene were identified. Comparison of 940-bp sequence 5′ to the putative translation start site reveals a highly conserved region (78.8% nucleotide identity) between *Hdh* and the HD gene from nucleotide −56 to −206 (of *Hdh*). Neither *Hdh* nor the HD gene have typical TATA or CCAAT elements, but both show one putative AP2 binding site and numerous potential Sp1 binding sites. The high sequence identity between *Hdh* and the HD gene for approximately 200 bp 5′ to the putative translation start site indicates that these sequences may play a role in regulating expression of the Huntington disease gene. © 1995 Academic Press, Inc.

INTRODUCTION

Huntington disease (HD) is an autosomal dominant neurodegenerative disorder characterized by involuntary movements, psychological disturbance, and cognitive decline that usually manifests in mid-life (Hayden, 1981; Harper, 1991). Recently, a novel gene containing a CAG trinucleotide repeat that is expanded on HD chromosomes was identified (HDCRG, 1993). This gene encodes two messenger RNAs that are widely expressed in different tissues but with varying abundance (Lin *et al.*, 1993; Li *et al.*, 1993; Strong *et al.*, 1993; Ambrose *et al.*, 1994). We and others have cloned the mouse homologue (*Hdh*) of the human HD gene (Lin *et al.*, 1994; Barnes *et al.*, 1994) and mapped it to chromosome 5 within a region of conserved synteny with human 4p16.3 (Nasir *et al.*, 1994). However, little is known regarding either the function of the HD gene product or the regulation of expression of the HD gene.

As a first step to further our understanding of the organization and regulated expression of the HD gene, we have cloned the genomic regions containing the first five exons, including exon 1, which contains the CAG repeat from mouse, and determined their genomic organization. We have conducted a detailed structural comparison of 5′ upstream sequences, including the putative promoter region between the human and the mouse HD genes.

Both the CAG and the CCG repeats immediately following the CAG repeat in the HD gene have been shown to be polymorphic in the general population (Andrew *et al.*, 1994; Kremer *et al.*, 1994; Rubinzstein *et al.*, 1994). The CCG repeat in the HD gene varies from 6 to 12, and the CAG repeat length varies from 9 to 35 on normal chromosomes (Kremer *et al.*, 1994). We have previously shown that the CCG repeat between nucleotide positions 211 and 223 of *Hdh* is polymorphic between two different mouse strains (Lin *et al.*, 1994). There are three CCG repeats in the ICR outbred strain (Lin *et al.*, 1994) and in wild mouse *Mus spretus* (Barnes *et al.*, 1994), but four CCG repeats in 129J, C57BL/6J (Lin *et al.*, 1994), PCC4, and CBA strains (Barnes *et al.*, 1994). In contrast to the murine CCG repeat, the adjacent $(CAG)_2CAA(CAG)_4$ is not polymorphic in five strains of mice (129J, PCC4, CBA, C57BL/6J, and ICR outbred) (Lin *et al.*, 1994; Barnes *et al.*, 1994).

Comparison of about 940-bp sequences 5′ of the putative translation start site has, however, identified a highly conserved region between the *Hdh* and the HD genes from −56 to −206 with a nucleotide identity of 78.81%, suggesting that these sequences in the 5′ flanking regions of both the *Hdh* and the HD genes

[1] To whom correspondence should be addressed at the Department of Medical Genetics, University of British Columbia 416-2125 East Mall, NCE Building Vancouver, B.C. Canada V6T 1Z4. Telephone: (604) 822-9240. Fax: (604) 822-9238.

0888-7543/95 $6.00

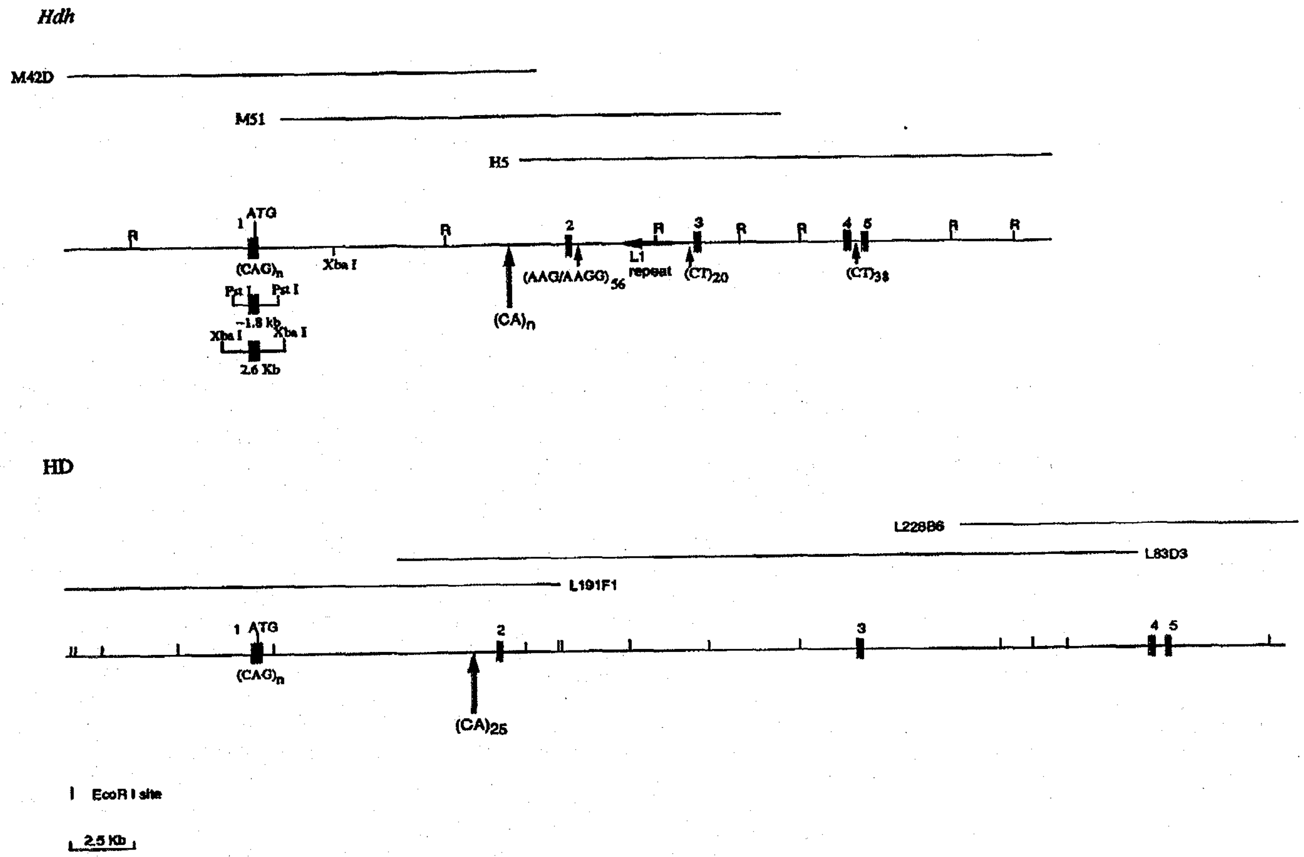


FIG. 1. Schematic map showing the comparative genomic organization of the 5′ region of *Hdh* and the HD gene. Black boxes represent exons. M42d, M51, and H5 are phages isolated from a mouse genomic DNA library, while L191F1, L83D3, and L228B6 (HDCRG, 1993) are cosmids from the chromosome 4-specific cosmid genomic library (courtesy of Lawrence Livermore National Laboratory). Microsatellite repeats are indicated by arrows.

may be important in the regulated expression of these genes.

MATERIALS AND METHODS

Isolation of genomic clones for the 5′ region of the Hdh and HD gene. A total phage genomic library of mouse strain 129J was plated at high density (200,000 pfu/24 × 24 cm^2 bioassay dishes) onto NZY media. Two sets of replica filters were made from each plate using Hybond-N$^+$ nylon filters (Amersham). The filters were immersed in denaturing solution (1.5 *M* NaCl, 0.5 *M* NaOH) for 30 s, in neutralization solution [1.5 *M* NaCl, 0.5 *M* Tris–HCl (pH 8.0)] for 30 s, in 2× SSC for 30 s and baked at 80°C for 2 h.

After secondary and tertiary screening, the positive plaques were picked. DNA from these positive phage were extracted, digested with different enzymes, and transferred onto Hybond-N$^+$ nylon filters (Amersham). Genomic fragments containing exons were identified by hybridization with *Hdh* cDNA or primers designed from the sequences of the cDNA (Lin *et al.*, 1994).

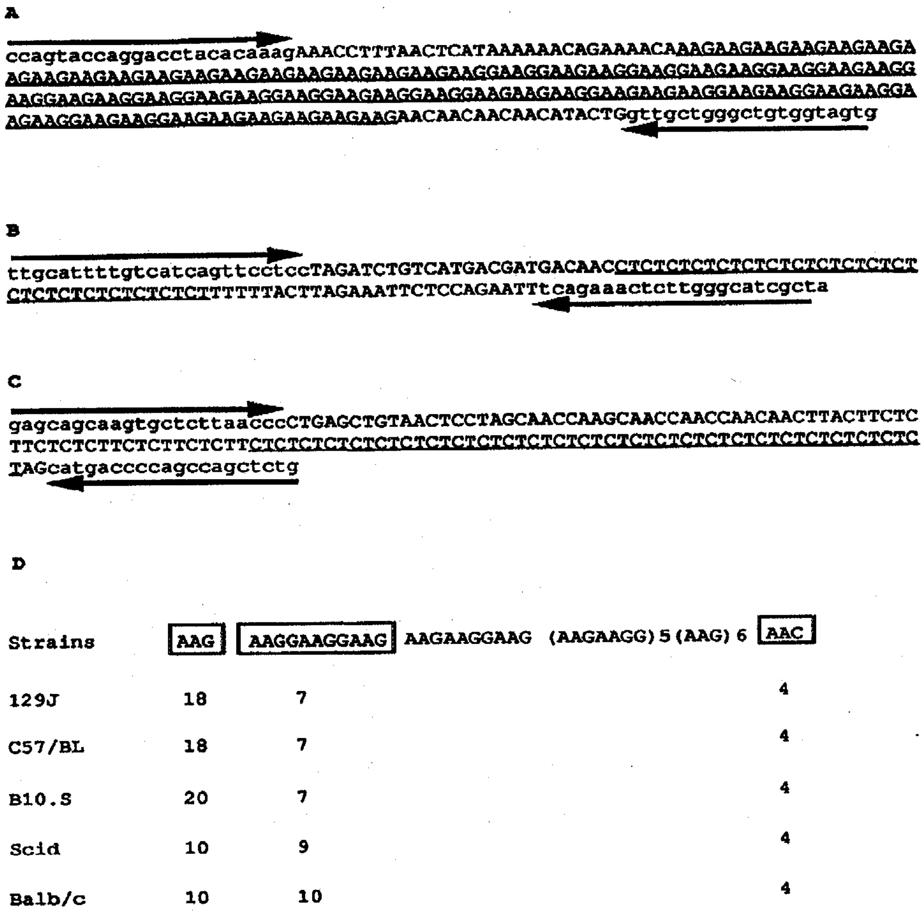
Human cosmids L191F1, L83D3, and L228B6 (HDCRG 1993) encompassing the 5′ genomic region of HD gene were picked from a gridded human chromosome 4-specific library (cell source: UV20 HL21-27, hamster–human hybrid cell lines containing human chromosomes 4, 8, and 21), and genomic cosmid blots were made after DNA from these cosmids was digested with *Eco*RI, *Hin*dIII, and *Pst*I. The filters were hybridized with primers designed from the human HD cDNA to map the exons.

Prehybridization and hybridization. Prehybridization and hybridization were performed in Church buffer (0.5 *M* sodium phosphate buffer, pH 7.2, 7% SDS, and 1 m*M* EDTA) at 65°C (Church and Gilbert, 1984). After hybridization, filters were washed gradually to a final stringency of 1× SSPE (0.18 *M* NaCl, 0.01 *M* NaH_2PO_4, 1 m*M* EDTA, pH 7.7) and 0.1% SDS at 65°C for 20 min. Autoradiography was carried out for 12–24 h at −70°C. Positive clones were purified following secondary and tertiary screening.

DNA sequencing and analysis. Plasmid DNA was prepared using a plasmid DNA preparation column (Qiagen). Automated sequencing was performed using the ABI373A sequencer. Manual dideoxy sequencing was performed using the Sequenase kit (US Biochemicals). Sequencing and PCR primers were synthesized with a PCR Mate 391 DNA synthesizer (Applied Biosystems, Inc.).

Sequence data were entered into a Sun IPX Workstation and analyzed with the Staden sequence analysis package (Dear and Staden, 1991). The commercial sequence analysis program MacVector (International Biologies, Inc.) and GeneWorks (IntelliGenetics Inc.) were also used for sequence analyses and for identification of potential *cis*-regulatory elements within the mouse and human HD putative promoters.

Assessment of polymorphisms. Further sequence analysis of the mouse gene has revealed three additional microsatellite repeats in



A

```
ccagtaccaggacctacacaaagAAACCTTTAACTCATAAAAAACAGAAAACAAAGAAGAAGAAGAAGAAGA
AGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGGAAGGAAGAAGGAAGGAAGAAGGAAGGAAGAAGG
AAGGAAGAAGGAAGGAAGAAGGAAGGAAGAAGGAAGGAAGAAGAAGGAAGAAGAAGGAAGAAGGAAGAAGGA
AGAAGGAAGAAGGAAGAAGAAGAAGAAGAAGAACAACAACAACATACTGgttgctgggctgtggtagtg
```

B

```
ttgcatttgtcatcagttcctccTAGATCTGTCATGACGATGACAACCTCTCTCTCTCTCTCTCTCTCTCT
CTCTCTCTCTCTCTCTTTTTTACTTAGAAATTCTCCAGAATTtcagaaactcttgggcatcgcta
```

C

```
gagcagcaagtgctcttaacccCTGAGCTGTAACTCCTAGCAACCAAGCAACCAACCAACAACTTACTTCTC
TTCTCTCTTCTCTTCTCTTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTC
TAGcatgaccccagccagctctg
```

D

Strains	AAG	AAGGAAGGAAG	AAGAAGGAAG	(AAGAAGG)5(AAG)6	AAC
129J	18	7			4
C57/BL	18	7			4
B10.S	20	7			4
Scid	10	9			4
Balb/c	10	10			4

FIG. 2. (A) Sequence of the $(AAG/AAGG)_{56}$ repeats. The PCR primer sequence for amplifying the repeats are in lowercase and indicated by arrows, while the repeat sequences themselves are underlined. (B) Sequence of the $(CT)_{20}$ repeats. (C) Sequence of the $(CT)_{27}$ repeats. (D) Polymorphism of the AAG/AAGG repeat in different mouse strains. The repeat sequences in the box are polymorphic. The numbers under the boxed repeat sequence indicate the number of the corresponding repeat. The polymorphism is polar and seen in the 5′ portion of the AAG/AAGG repeat but absent in the 3′ portion of the repeat in five different strains of mice (129J, C57BL/6J, B10.S, Scid, and Balb/c).

the introns of *Hdh*: the $(AAG/AAGG)_{56}$ repeat and the $(CT)_{20}$ repeat in intron 2 and the $(CT)_{27}$ repeat in intron 4 (Fig. 1). The sequences of these repeats and their flanking regions have been deposited in GenBank under Accession Nos. L34021 through L34023.

We have designed PCR primers to assess for length polymorphisms of these repeats in several different strains of laboratory mice (*Mus musculus*) including 129J, Nude, Scid, Balb/c, B10.S, CBA, C57BL/6J, and a wild species of mouse *Mus spretus*. The primers are 5′-CCAGTACCAGGACCTACACAAAG-3′ (forward) and 5′-CACTACCACAGCCCAGCAAC-3′ (reverse) for the $(AAG/AAGG)_{56}$ repeats; 5′-TTGCATTTTGTCATCAGTTCCTCC-3′ (forward) and 5′-TAGCGATGCCCAAGAGTTTCTGA-3′ (reverse) for the $(CT)_{20}$ repeats; 5′-GAGCAGCAAGTGCTCTTAACCC-3′ (forward) and 5′-CAGAGCTGGCTGGGGTCATG-3′ for the $(CT)_{27}$ repeats. The sequence of the repeats is shown in Fig. 2.

The expected sizes of the PCR products for 129J strain of mouse (Table 1) are 285 bp for the $(AAG/AAGG)_{56}$ repeats, 137 bp for the $(CT)_{20}$ repeats, and 167 bp for the $(CT)_{27}$ repeats. A total of 4 pmol of one primer of each PCR primer set was end-labeled separately with [γ-^{32}P]ATP and T4 polynucleotide kinase in a 25-μl reaction. In each PCR, 0.5 μl of the labeled primer was included. Each PCR reaction was performed with 0.1 μg of genomic DNA, 0.4 pmol of each primer, 0.12 m*M* dNTPs, 1.5 m*M* $MgCl_2$, and 1× PCR buffer (50 m*M* KCl, 10 m*M* Tris–HCl, pH 9.0, at 25°C and 0.1% Triton X-100) for 35 cycles. Thermal cycling conditions were 94°C 30 s, 56°C 30 s, and 72°C 30 s for the $(AAG/AAGG)_{56}$ repeats; 94°C 30 s, 62°C 30 s, and 72°C 30 s for the $(CT)_{20}$ repeats; 94°C 30 s, 64°C 30 s, and 72°C 30 s for the $(CT)_{27}$ repeats. After amplification, 4 μl of each of the PCR products was mixed with 4 μl of the formamide loading dye, denatured at 80°C for 2 min, loaded onto a 6% polyacrylamide gel, and run together with a sequencing reaction as size marker. The gel was then dried and autoradiographed.

RESULTS

Comparison of the 5′ Upstream Sequences of the Hdh and HD Genes

We have subcloned a 4.1-kb *Eco*RI fragment, containing both the putative promoter region and exon 1 of the HD gene, and a 2.6-kb *Xba*I fragment containing the putative *Hdh* promoter region. We have sequenced

TABLE 1

CT Repeat Polymorphisms in the *Hdh* Gene in Different Strains of Mice

	Number of CT repeats	
Strains	Intron 2	Intron 4
129J	20	27
Nude	22	25
Scid	21	27
Balb/c	21	27
B10.S	20	26
CBA	21	26
C57/BL	20	26
Mus spretus	18	12

the entire 4.1-kb *Eco*RI fragment including 3580 bp upstream of the putative ATG translation start codon of the HD gene (GenBank Accession No. L34020). Similarly, we sequenced the promoter region of *Hdh* (GenBank Accession No. L34008) including about 930 bp upstream of the putative ATG translational start site.

Sequence alignment of the 5′ region upstream of the putative ATG codon of *Hdh* and HD is shown in Fig. 3. This alignment reveals the existence of a highly conserved region between *Hdh* and the HD gene from −56 to −206 (numbered after *Hdh*) with a nucleotide identity of 78.81%. Identity is less significant between −55 and +1 (translation start site) (54% nucleotide identity), while upstream of sequences −207 to −930 of *Hdh* is only about 50% nucleotide identity. Within the conserved region, one cAMP responsive element (TGACGTCA) (Andrisani *et al.,* 1988) is present at position −180 to −174 in *Hdh* but not in the HD gene.

Sequence analyses of the 5′ region of the HD gene have also revealed the existence of two 20-bp pairs direct repeats (GGCCCCGCCTCCGCCGGCGC) at position −212 to −193 and −192 to −173 with only one mismatch (Figs. 3 and 4). However, these repeats are not present in the 5′ upstream sequences of *Hdh*. We have also identified two direct 17-bp repeats (CCACGCCCCCCGCATCG) at position −516 to −500 and −499 to −483 in the HD gene (Figs. 3 and 4). These two 17-bp repeats were flanked by CCACGCC repeats that are identical to the first 7 bp of the 17-bp repeats.

Assessment for transcriptional protein-binding motifs using a commercially available software package (MacVector 3.5) revealed a conserved AP2 (CCGCAGGC) site at position −248 to −240 in the HD gene and −270 to −262 in *Hdh* (Fig. 3). There are also 5 potential Sp1 binding sites (GGGCGG) at position −299 to −293, −318 to −312, −374 to −368, −379 to −373, and −427 to −421 in *Hdh* and 11 potential Sp1 sites in the HD gene at position −15 to −9, −284 to −278, −453 to −447, −541 to −535, −571 to −565, −587 to −581, −592 to −586, −638 to −632, −643 to −637, −654 to −648, and −706 to −700. However, only 1 of these Sp1 sites is conserved (Fig. 3).

Two tandem head to tail *Alu* repeat sequences were identified in the HD gene, both in the opposite direction to the transcription of HD gene. One, starting at −2099, is a full-length *Alu* repeat belonging to the *Alu*-Sx subfamily (flanked by CTGGGAACTT direct repeats) as detected after searching the *Alu* databases with PYTHIA server (Milosavljevic and Jurka, 1993; Jurka and Milosavljevic, 1991; Hutchinson *et al.,* 1993). The other *Alu,* starting at −1723 nucleotide position, is a truncated (half) *Alu* repeat retaining only the 3′ end of the *Alu* sequences and belongs to the *Alu*-J subfamily (Jurka and Milosavljevic, 1991).

Comparative Genomic Organization of the First 5 Exons of the HD Gene and Hdh

A phage contig encompassing the first 5 exons of *Hdh* was constructed (Fig. 1). Phage H5 was initially isolated by hybridization with a *Hdh* cDNA probe, MHD2 (Lin *et al.,* 1994). A 3.5-kb *Sph*I fragment near the 5′ end of phage H5 was used to initiate a chromosome walk to extend the contig. A second positive phage clone (M51) was isolated, but this clone also did not contain the CAG repeat. Thereafter, a PCR product spanning nucleotides 1 to 123 of the *Hdh* cDNA was used to screen the genomic library, and a third phage clone (M42A-D) was isolated and shown to contain the CAG exon by hybridization to the CAG repeat.

Exon-containing fragments were identified by hybridization with primers derived from *Hdh* cDNA sequences. These fragments were then subcloned and sequenced (GenBank Accession Nos. L34008 and L34021 to L34024). The intron/exon junction sequences of the first 5 exons of the *Hdh* all conform to the GT–AG rules (Padgett *et al.,* 1986) (Table 2). The first 5 nucleotides of the 5′ donor splice sites are identical between *Hdh* and HD, while only the AG in the 3′ acceptor splice sites are identical (Table 2).

Exons 2 to 5 of *Hdh* are identical in size to the HD gene, with donor and acceptor splice sites at the same homologous positions in the HD cDNA sequence (Ambrose *et al.,* 1994). Furthermore, these exons are highly conserved with nucleotide identity of 93, 91, 92, and 95% for exons 2, 3, 4, and 5, respectively. However, exons 1 of the *Hdh* and the HD gene are different in size and divergent in sequence identity because of differences not only in the length of the CAG repeat length (7 in mouse and a mean of 18 in human) but also in that the HD gene has 9 extra CCN (N = G, C, or A) repeats and other nucleotide sequences (Lin *et al.,* 1994).

Intron sizes differ markedly between the HD gene and *Hdh* (Table 2 and Fig. 1). Furthermore, the introns have significant differences in sequence. Comparison of the intron sequences of *Hdh* with the intron sequences of the HD genes including 247 bp of the 5′ end of intron 1 and other limited (60 bp) amounts of intron sequences of the HD gene available in GenBank (GenBank Accession Nos. L27350–L27354) (Ambrose *et al.,* 1994) reveals only approximately 50% nucleotide identity excluding the splice junction signals—GT–AG. We have also identified an L1 repeat truncated at the 5′ end in intron 2 of the *Hdh* oriented in the opposite direction to the direction of transcription of *Hdh* (Fig. 1).

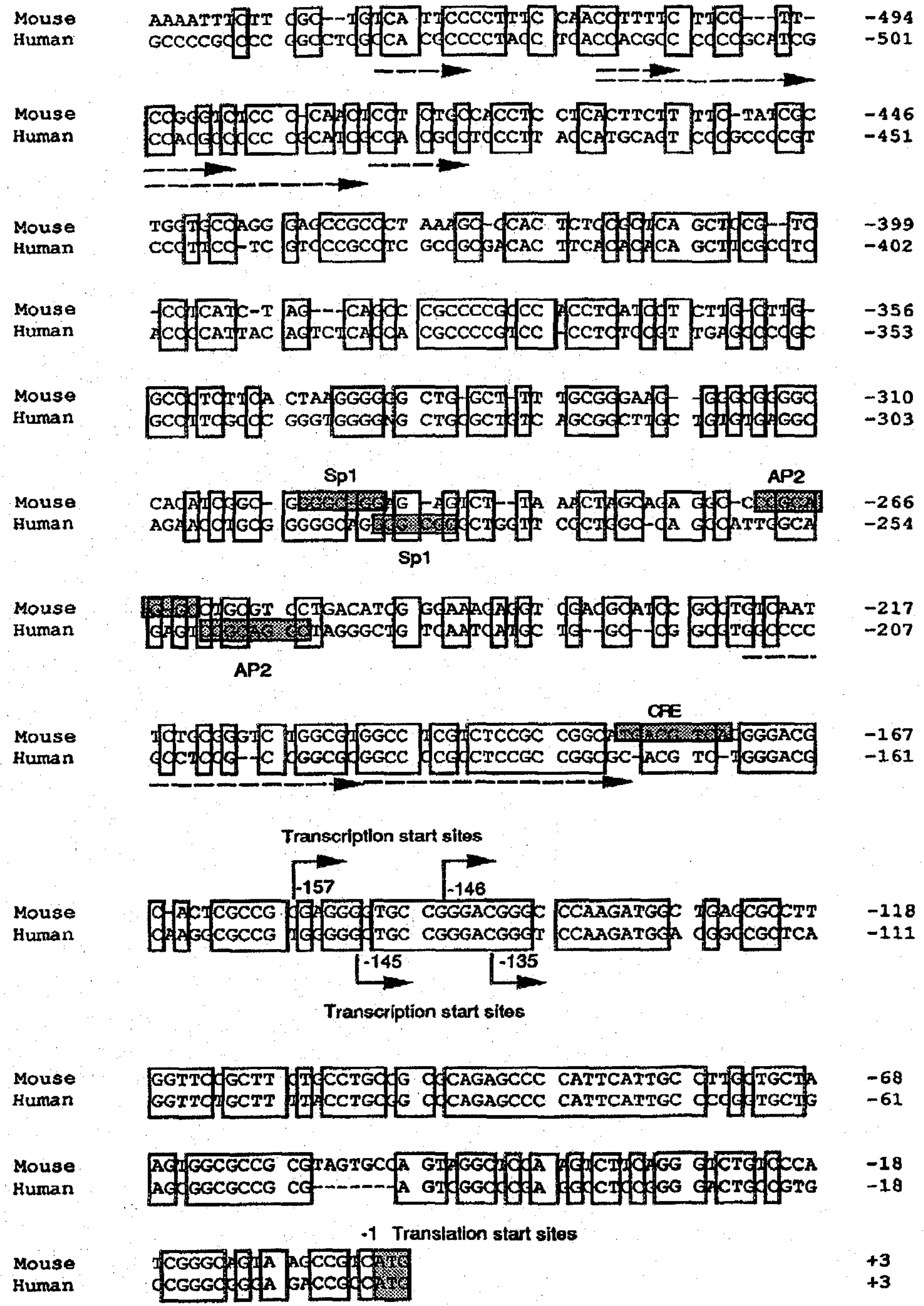

Mouse AAAATTTCTT CGC--TGTCA TTCCCCTTTC CAACCTTTTC TTCC---TT- -494
Human GCCCCGCCCC GGCCTCGCCA CGCCCCTACC TCACCACGCC CCCCGCATCG -501
Mouse CCGGGTCTCC C-CAACTCCT CTGCCACCTC CTCACTTCTT TTC-TATCGC -446
Human CCACGCCCCC CGCATCGCCA CGCCTCCCTT ACCATGCAGT CCCGCCCCGT -451
Mouse TGGTGCCAGG GAGCCGCCCT AAAGC-CCAC TCTCGGCTCA GCTCCG--TC -399
Human CCCTTCC-TC GTCCCGCCTC GCCGCGACAC TTCACACACA GCTTCGCCTC -402
Mouse -CCTCATC-T AG---CAGCC CGCCCCGCCC ACCTCATCCT CTTG-CTTG- -356
Human ACCCCATTAC AGTCTCACCA CGCCCCGTCC -CCTCTCCGT TGAGCCCCGC -353
Mouse GCCCTCTTCA CTAAGGGGGG CTG-GCT-TT TGCGGGAAG- -GGCGGCGGC -310
Human GCCTTCGCCC GGGTGGGGNG CTGCGCTCTC AGCGGCTTGC TGTGTGAGGC -303
Sp1
AP2
Mouse CACATCGGC- GGGGCGGGAG -AGTCT--TA AACTAGCAGA GGC-CGCCCC -266
Human AGAACCTGCG GGGGCAGCCG CCGCCTGGTT CCCTGGC-CA GCCATTGGCA -254
Sp1
Mouse GCGCCTGCGT CCTGACATCG GCAAAGAGGT CGACGCATCC GCCTGTCAAT -217
Human GAGTGCCCAG CCTAGGGCTG TCAATCATGC TG--GC--CG GCCTGGCCCC -207
AP2
CRE
Mouse TCTGCGGTCC TGGCGTGGCC TCGTCTCCGC CGGCATGACG TCAGGGACG -167
Human GCCTCCG--C CGGCGCGGCC CCGCCTCCGC CGGCGC-ACG TC-TGGGACG -161
Transcription start sites
-157
-146
Mouse C-ACTCGCCG CGAGGGCTGC CGGGACGGGC CCAAGATGGC TGAGCGCCTT -118
Human CAAGGCGCCG TGCGGGCTGC CGGGACGGGT CCAAGATGGA CGGGCGCTCA -111
-145
-135
Transcription start sites
Mouse GGTTCCGCTT CTGCCTGCCG CGCAGAGCCC CATTCATTGC CTTGCTGCTA -68
Human GGTTCTGCTT TTACCTGCGG CGCAGAGCCC CATTCATTGC CCCGGTGCTG -61
Mouse AGTGGCGCCG CGTAGTGCCA GTAGGCTCCA AGTCTTCAGG GTCTGTCCCA -18
Human AGCGGCGCCG CG-------A GTCGGCCCGA GGCCTCCCGG GACTGCCGTG -18
-1 Translation start sites
Mouse TCGGGCAGTA AGCCGTCATG +3
Human CCGGGCGGGA GACCGCCATG +3

FIG. 3. Sequence alignment of the 5′ region upstream of the putative translation start codon between *z* (GenBank Accession No. L34008) and HD (GenBank Accession No. L34020). Identical sequences are boxed. The alignment was performed with the DNA sequence analysis software GeneWorks 2.5 (IntelliGenetics). AP2 and SP1 sites and cAMP-responsive element binding sites (CRE) are shown in the gray boxes.

Comparison of Three Microsatellite Repeats in the Introns of the Hdh and HD Genes

We have identified three microsatellite repeats in *Hdh* (Fig. 1). They are $(AAG/AAGG)_{56}$ repeats and $(CT)_{20}$ repeats in intron 2 and $(CT)_{27}$ repeats in intron 4. The AAG repeat is interrupted 20 times by AAGG following 18 perfect AAG repeats. The $(CT)_{27}$ repeats are preceded by a $(CTTCT)_2(CTCTT)_3$ repeat. These mi-

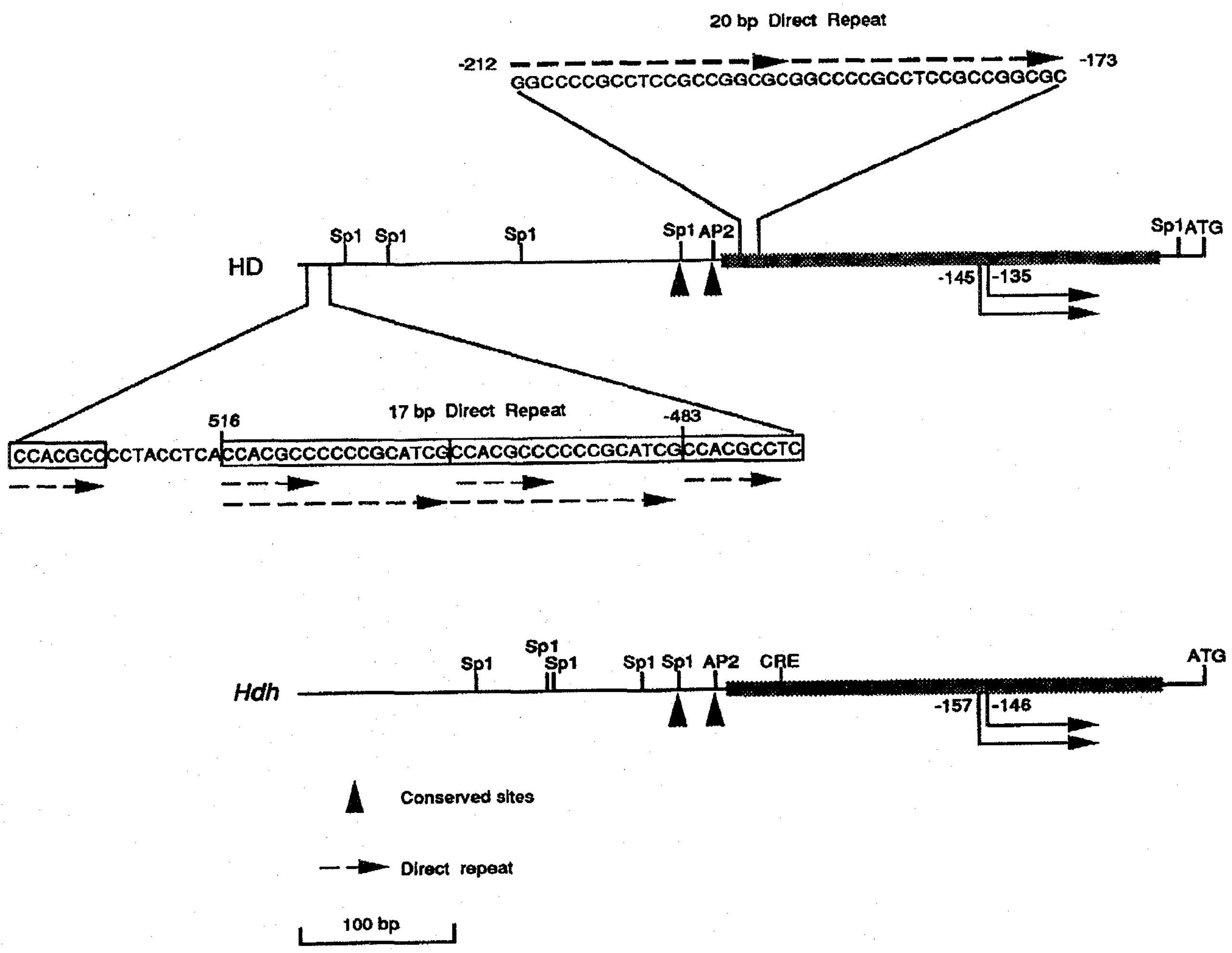


FIG. 4. Schematic map showing the relative positions of the putative transcription factor binding sites in the 550 bp upstream of the putative translation start site of *Hdh* and HD. The published *Hdh* and human HD cDNAs are represented by arrows. The direct repeat sequences found in the HD gene are indicated by dashed arrows. Conserved Sp1 and Ap2 sites are indicated by arrowheads. The region of conservation is indicated by a gray box. The putative translation start site (ATG) is designated −1.

crosatellite repeats are not present in the homologous region of the HD gene, as the oligonucleotide primers corresponding to those repeats failed to hybridize to the DNAs from cosmids spanning this genomic region. A total of four different alleles were identified both for the $(CT)_{20}$ repeats and for the $(CT)_{27}$ repeats (Table 1). The allele sizes of all three microsatellite repeats are the smallest for the wild species of mouse—*Mus spretus*.

We have also sequenced the $(AAG/AAGG)_{56}$ repeat from six different inbred mouse strains (Fig. 2D). It is evident that the polymorphism of the $(AAG/AAGG)_{56}$ repeats is polar arising from variation within the first 18 consecutive AAG repeats. The first consecutive run of AAG repeats varies from 20 in B10.S strains to 18 in C57BL/6J and 129J strains and 10 in scid and Balb/c strains. An adjacent AAGGAAGGAAG repeat acts as a longer repeat unit and varies from 10 repeat units in balb/c to 9 in scid and 7 in C57BL/6J, B10.S, and 129J strains. The rest of the sequences [AAGAAGGAAG$(AAGAAGG)_5(AAG)_6$] are not polymorphic in 129J, B10.S, scid, C57BL/6J, and Balb/c strains of mice.

We have also identified a polymorphic CA repeat in intron 1 of the HD gene. Its repeat number is 25 on cosmid 191F1. Analysis of 44 normal human chromosomes of Western European origin revealed that the CA_n is highly polymorphic with 9 distinct allele sizes including CA_{23} (4.5%), CA_{24} (13.6%), CA_{25} (6.8%), CA_{26} (20.5%), CA_{27} (15.9%), CA_{30} (11.4%), CA_{31} (6.6%), and CA_{32} (4.1%). This CA repeat is also found in intron 1 of *Hdh* as detected by hybridization with a $(CA)_{10}$ oligonucleotide.

TABLE 2

Comparison of the Exon and Intron Size and Splice Site Sequences of the First 5 Exons of the Mouse and Human HD Genes[a]

Exon	Gene	Exon size (bp)	5′ donor site	Intron	Intron size (kb)	3′ acceptor site
1	*Hdh*		GTGAGTCCGGGCGCCGCAGCTC	1	~15	TTTTCCTCTTGTTTTTTTGTAG
1	HD		GTGAGTTTGGGCCCGCTGCAGC	1	~10	TCCTTCTTTTTTTTATTTTTAG
2	*Hdh*	84	GTAATTGGCTTTTTAAAAAAAA	2	~7	TCTCTCTCTCTTTTTTACTTAG
2	HD	84	GTAATTGCACTTTGAACTGTCT	2	~15	TTTCTCTTCTTTTTTTGCTTAG
3	*Hdh*	121	GTAAGCGCCCCATAATGATGAT	3	~5	AGTCTCTTCTATTTCTTTGCAG
3	HD	121	GTAAGAACCGTGTGGATGATGT	3	~7	AATCTCTTGTGATTTGTTGTAG
4	*Hdh*	60	GTGGGTGTTTGCTCTGCATTAT	4	~0.5	ATCACTTGTTAACTCCACTTAG
4	HD	60	GTGGGCCTTGCTTTTCTTTTTT	4	~0.5	AACCCTCATTGCACCCCCTCAG
5	*Hdh*	80	GTAAGTTGTACCTCTGTATTATTTTTAAGA			
5	HD	80	GTAAGTTGTACACTCTGGATGTTGGTTTTT			

[a] Donor and acceptor splice sites of HD are from Ambrose *et al.* (1994).

DISCUSSION

The high GC content and lack of both typical TATA and CAAT *cis*-elements in the 5′ flanking regions in *Hdh* and HD suggest that they are "housekeeping" genes. Indeed, *Hdh* and HD are both abundantly expressed in different tissues (Lin *et al.*, 1993, 1994; Li *et al.*, 1993; Strong *et al.*, 1993; Ambrose *et al.*, 1994). The presence of a highly conserved region in the 5′ flanking region between *Hdh* and HD from −56 to −206 (78.81% nucleotide identity) suggests that these regions may play a critical role in regulating expression of the HD gene. In support of this, preliminary mapping of the transcription initiation sites of both *Hdh* and HD show two major transcription initiation sites at −157 and −146 in *Hdh* and −135 and −145 in HD (data not shown).

The presence of the 17-bp direct repeats (CCACGCCCCCCGCATCG) and the two 20-bp pair perfect repeats (GGCCCCGCCTCCGCCGGCGC) in the human HD gene may serve as unique binding sites for *trans*-acting factors that may either direct transcriptional initiation or enhance expression of the HD gene. Many direct repeats described to date are located within well-defined promoters. For example, the Chinese hamster ovary dihydrofolate reductase (*dhfr*) and human low-density lipoprotein (LDL) receptor promoters contain unique 29- and 16-bp direct repeat sequences, respectively (Mitchell *et al.*, 1985; Südhof *et al.*, 1987), necessary for both transcriptional activation and regulated expression. The absence of these repetitive elements, however, in the 5′ flanking region of *Hdh* would suggest that the expression of these two genes is regulated, in part, by different *cis*- and/or *trans*-regulatory elements. There are four 29-bp direct repeats found in the hamster and mouse *dhfr* promoter, compared to one in the human *dhfr* homologue, with no obvious divergent function of the protein (Mitchell *et al.*, 1986). In contrast, the 16-bp imperfect direct repeats found in the promoter of the LDL receptor gene have remained highly conserved in the hamster, mouse, and human, both in nucleotide sequence and relative position. These three direct repeats have been shown to serve distinct functions, two of which are responsible for binding the *trans*-regulatory protein Sp1, whereas one acts as a sterol-responsive element (Bishop, 1992; Südhof *et al.*, 1987). The conservation of these repeat motifs in the LDL receptor promoter alludes to their importance in controlling transcription of this gene. In contrast, the lack of conservation of these repeats within the *Hdh* and the HD gene putative promoter regions may represent the evolution of genes with different patterns of regulation.

The polar variation of the AAG/AAGG repeat in intron 2 of *Hdh* is similar to that observed in other human trinucleotide (Kunst and Warren, 1994) and minisatellite (Armour *et al.*, 1993) repeats. In the AAG/AAGG repeat reported here, the 5′ portion consists of the AAG repeat without interruption and thus may be more unstable, whereas the 3′ portion consists of interrupted repeats and would therefore be more stable. It has been suggested that polar variation at repeat loci might be a general phenomenon in the human genome and implies that mutation within these repeats is regulated at least in part by the nature of the sequence itself (Armour *et al.*, 1993). The polar variation of this repeat in the *Hdh* suggests that this phenomenon described in human genes may be more widespread and is also evident in nonhuman genomes.

An important question is why the murine CAG repeat is not polymorphic and is considerably smaller than the CAG repeat size in the human gene. In contrast to the CAG repeat in the HD gene, the CAG repeat in *Hdh* is cryptic (interrupted by a CAA repeat after two CAG repeats) (CAGCAGCAACAGCAGCAGCAG). Kunst and Warren (1994) have recently demonstrated that an AGG repeat interruption in the CGG repeat in the FRAXA locus confers some degree of stability for the CGG repeat in the gene associated with fragile X syndrome. Similarly, for spinocerebellar ataxia type 1 (SCA1), an uninterrupted $(CAG)_n$ repeat configuration is seen on unstable alleles in the gene, whereas when CAG is interrupted by CAT the CAG repeat is more stable (Chung *et al.*, 1993). This suggests that an important factor that may be contributing to the stability of the CAG repeat in *Hdh* is its cryptic nature, which may at least in part account for the fact that the CAG repeat number in *Hdh* (7 CAG repeats) is significantly less than the CAG repeat number (mean = 18) in the HD gene. This may also explain why no naturally occurring murine model for HD has been identified; it may not exist.

The $(CT)_{20}$, $(CT)_{27}$, and $(AAG/AAGG)_{56}$ repeats identified in the intron of *Hdh* were not found in the homologous region of the HD gene. In contrast, the (CA) repeat was identified in intron 1 of both the *Hdh* and the HD genes. Stallings (1994) compared 10 different trinucleotide repeats from which the corresponding homologous region sequences were available between human and rodent and found no conservation of trinucleotide in similar regions. Also, Stallings *et al.* (1991) compared 17 $(GT)_n$ repeats between rodents and humans and found that only 4 of these GT repeats (23.5%) are localized in the same map location in both species. It is therefore not surprising that the AAG/AAGG repeat and the two CT repeats in the introns of *Hdh* are not present in the HD gene, as they could have arisen in the mouse gene after the divergence of primates and rodents. The $(CT)_{27}$ repeat will be convenient for use in mapping experiments because the repeat length differences between *Mus musculus* and *Mus spretus* are on average about 30 and 15 bp, respectively, which make them detectable in a regular agarose gel and eliminates the need to use hot-PCR.

In summary, we have performed structural analyses of the 5′ region, including the promoter, between *Hdh* and HD. Within the promoter, there is one markedly conserved region (−56 to −206) between mouse and human genes that will now allow functional analysis of this region to determine its role in the regulation of this gene. The absence of conservation of putative transcription binding motifs between HD and *Hdh* suggests differences in regulation of these genes in mouse and human.

ACKNOWLEDGMENTS

This work was support by grants from the Medical Research Council of Canada, the Canadian Genetic Disease Network, and the Huntington Disease Society of Canada. M.R.H. is an established investigator with the B.C. Children's Hospital. J.N. is a recipient of the Bluma–Tischler Award.

REFERENCES

Ambrose, C. M., Duyao, M. P., Barnes, G., Bates, G. P., Lin, C. S., Srinidhi, J., Baxendale, S., Hummerich, H., Lehrach, H., Altherr, M., Wasmuth, J., Buckler, A., Church, D., Houseman, D., Berks, M., Micklem, G., Durbin, R., Dodge, A., Read, A., Gusella, J., and MacDonald, M. E. (1994). Structure and expression of the Huntington's disease gene: Evidence against simple inactivation due to an expanded CAG repeat. *Somatic Cell Mol. Genet.* **20:** 27–38.

Andrew, S. E., Goldberg, Y. P., Theilmann, J., Zeisler, J., and Hayden, M. R. (1994). A CCG repeat polymorphism adjacent to the CAG repeat in the Huntington disease gene: Implications for diagnostic accuracy and predictive testing. *Hum. Mol. Genet.* **3:** 65–67.

Andrisani, O. M., Pot, D. A., Zhu, Z., and Dixon, J. E. (1988). Three sequence-specific DNA–protein complexes are formed with the same promoter element essential for expression of the rat somatostatin gene. *Mol. Cell. Biol.* **8:** 1947–1956.

Armour, J. A., Harris, P. C., and Jeffreys, A. J. (1993). Allelic diversity at minisatellite MS205 (D16S309): Evidence for polarized variability. *Hum. Mol. Genet.* **2:** 1137–1145.

Barnes, G. T., Duyao, M. P., Ambrose, C. M., McNeil, S., Perischetti, F., Srinidhi, J., Gusella, J. F., and MacDonald, M. E. (1994). Mouse Huntington's disease gene homolog (*Hdh*). *Somatic Cell Mol. Genet.* **20:** 87–97.

Bishop, R. W. (1992). Structure of the hamster low density lipoprotein receptor gene. *J. Lipid Res.* **33:** 549–557.

Chung, M.-Y., Ranum, L. P. W., Duvick, L., Servadio, A., Zoghbi, H. Y., and Orr, H. T. (1993). Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type 1. *Nature Genet.* **5:** 254–258.

Church, G. M., and Gilbert, W. (1984). Genomic sequencing. *Proc. Natl. Acad. Sci. USA* **81:** 1991–1995.

Dear, S., and Staden, R. (1991). A sequence assembly and editing program for efficient management of large projects. *Nucleic Acids Res.* **19:** 3907–3911.

Gwynn, B., Lyford, K. A., and Birkenmeier, E. H. (1990). Sequence conservation and structural organization of the glycerol-3-phosphate dehydrogenase promoter in mice and humans. *Mol. Cell. Biol.* **10:** 5244–5256.

Harper, P. S. (1991). "Huntington's Disease," W. B. Saunders, London.

Hayden, M. R. (1981). "Huntington Chorea," Springer-Verlag, New York.

Huntington Disease Collaborative Research Group (HDCRG) (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72:** 931–983.

Hutchinson, G. B., Andrew, S. E., McDonald, H., Goldberg, Y. P., Graham, R., Rommens, J. M., and Hayden, M. R. (1993). An Alu element retroposition in two families with Huntington disease defines a new active Alu subfamily. *Nucleic Acids Res.* **15:** 3379–3383.

Jurka, J., and Milosavljevic, A. (1991). Reconstruction and analysis of human Alu genes. *J. Mol. Evol.* **32:** 105–121.

Kremer, B., Goldberg, P., Andrew, S. E., Theilmann, J., Telenius, H., Zeisler, J., Squitieri, F., Lin, B., Bassett, A., Almqvist, E., Bird, T. D., and Hayden, M. R. (1994). A worldwide study of the Huntington's disease mutation: The sensitivity and specificity of measuring CAG repeats. *N. Engl. J. Med.* **330:** 1401–1406.

Kunst, C. B., and Warren, S. T. (1994). Cryptic and polar variation of the fragile X repeat could result in predisposing normal alleles. *Cell* **77:** 853–861.

Li, S. H., Schilling, G., Young, W. S., 3d., Li, X. J., Margolis, R. L., Stine, O. C., Wagster, M. V., Abbott, M. H., Franz, M. L., Ranen, N. G., *et al.* (1993). Huntington's disease gene (IT15) is widely expressed in human and rat tissues. *Neuron* **11:** 985–993.

Lin, B., Rommens, J. M., Graham, R. K., Kalchman, M., MacDonald, H., Nasir, J., Delaney, A., Goldberg, Y. P., and Hayden, M. R. (1993). Differential 3′ polyadenylation of the Huntington disease gene results in two mRNA species with variable tissue expression. *Hum. Mol. Genet.* **2:** 1541–1545.

Lin, B., Nasir, J., MacDonald, H., Hutchinson, G., Graham, R., Rommens, J. M., and Hayden, M. R. (1994). Sequence of the murine Huntington disease gene: Evidence for conservation, alternate splicing and polymorphism in a triplet (CCG) repeat. *Hum. Mol. Genet.* **3:** 85–92.

Ludwig, E. H., Levy-Wilson, B., Knott, T., Blackhart, B. D., and McCarthy, B. J. (1991). Comparative analysis of sequences at the

5′ end of the human and mouse apolipoprotein B gene. *DNA Cell Biol.* **10:** 329–338.

Milosavljevic, A., and Jurka, J. (1993). Discovery by minimal length encoding: A case study in molecular evolution. *Machine Learning J.* (Special Issue on Machine Discovery) **12:** 1, 2, 3.

Mitchell, P. J., Carothers, A. M., Han, J. H., Harding, J. D., Kas, E., Venolia, L., and Chasin, L. A. (1986). Multiple transcription start sites, DNase I-hypersensitive sites, and an opposite-strand exon in the 5′ region of the CHO dhfr gene. *Mol. Cell. Biol.* **6:** 425–440.

Nasir, J., Lin, B., Bucan, M., Koizumi, T., Nadeau, J., and Hayden, M. R. (1994). The murine homologue of the Huntington disease gene (*Hhd*) and the α-adducin gene (Add1) map to mouse chromosome 5 within a region of conserved synteny with human chromosome 4p16.3. *Genomics* **22:** 198–201.

Padgett, R. A., Grabowski, P. J., Konarska, M. M., Seiler, S., and Sharp, P. A. (1986). Splicing of messenger RNA precursors. *Annu. Rev. Biochem.* **55:** 1119–1150.

Rubinzstein, D. C., Amos, W., Leggo, J., Goodburn, S., Ramesar, R. S., Old, J., Bontrop, R., McMahon, R., Barton, D. F., and Ferguson-Smith, M. A. (1994). Mutational bias provides a model for the evolution of Huntington's disease and predicts a general increase in disease prevalence. *Nature Genet.* **7:** 525–530.

Stallings, R. L., Ford, A. F., Nelson, D., Torney, D. C., Hildebrand, C. E., and Moyzis, R. K. (1991). Evolution and distribution of $(GT)_n$ repetitive sequences in mammalian genomes. *Genomics* **10:** 807–815.

Stallings, R. L. (1994). Distribution of trinucleotide microsatellite in different categories of mammalian genomic sequence: Implications for human genetic diseases. *Genomics* **21:** 116–121.

Strong, T. V., Tagle, D. A., Vades, J. M., Elmer, L. W., Boehm, K., Swaroop, M., Kaatz, K. W., Collins, F. S., and Albin, R. L. (1993). Widespread expression of the human and rat Huntington's disease gene in brain and nonneural tissues. *Nature Genet.* **5:** 259–265.

Südhof, T. C., Van Der Westhuyzen, D. R., Goldstein, J. L., Brown, M. S., and Russell, D. W. (1987). Three direct repeats and a TATA-like sequence are required for regulation expression of the human low density lipoprotein receptor gene. *J. Biol. Chem.* **262:** 10773–10779.